(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 December 2022 (29.12.2022)

WIPO | PCT

(10) International Publication Number
WO 2022/266853 A1

(51) International Patent Classification:
G06K 9/00 (2022.01)

(21) International Application Number:
PCT/CN2021/101611

(22) International Filing Date:
22 June 2021 (22.06.2021)

(25) Filing Language: English

(26) Publication Language: English

(71) Applicant: **INTEL CORPORATION** [US/US]; 2200 Mission College Boulevard, Santa Clara, California 95054 (US).

(72) Inventors; and

(71) Applicants *(for BZ only)*: **GUO, Ping** [CN/CN]; Guangqu Jia Yuan 7#905, Dong Cheng District, Beijing 100022 (CN). **MANEPALLI, Sangeeta** [US/US]; 11635 E Bellflower Drive, Chandler, Arizona 85249 (US). **ZHANG, Lidan** [CN/CN]; 9F, Raycom InfoTech Park, A 2 Kexueyuan South Rd, Haidian, Beijing 100080 (CN). **WANG, Peng** [CN/CN]; 8F, Raycom, Park A, No. 2 Kexueyuan Sourth Road, Beijing 100192 (CN). **ZHANG, Yimin** [CN/CN]; 11-1507 Taiyangyuan Haidian District, Beijing 100098 (CN).

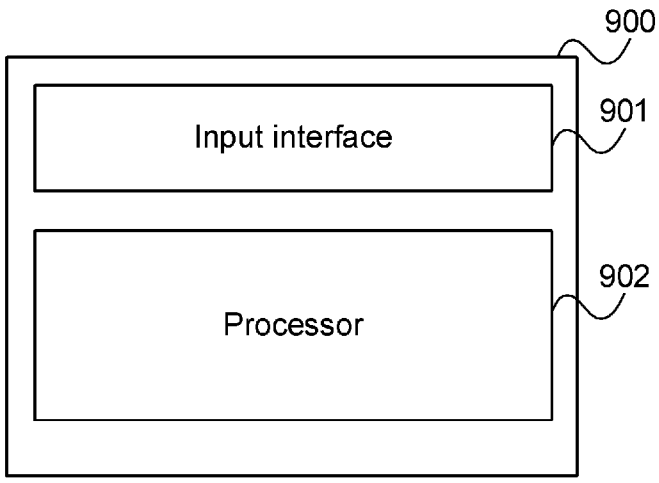
(74) Agent: **BEIJING EAST IP LTD.**; Suite 1601, Tower E2, The Towers, Oriental Plaza, No. 1, East Chang An Ave., Dongcheng District, Beijing 100738 (CN).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DJ, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, IT, JO, JP, KE, KG, KH, KN, KP, KR, KW, KZ, LA, LC, LK, LR, LS, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, WS, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, ST, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV,

(54) Title: METHODS AND DEVICES FOR GESTURE RECOGNITION

FIG 9



(57) Abstract: According to various examples, a gesture recognition device is described comprising an input interface configured to receive a sequence of images, each image showing a body part with which a gesture is performed from a viewpoint of a camera and a processor configured to generate a sequence of motion-compensated images from the sequence comprising generating a motion-compensated image for an image of the sequence by compensating the movement of the camera viewpoint from a reference camera viewpoint to the viewpoint from which the image shows the body part based on the image and a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence and estimate the gesture from the sequence of motion-compensated images.

WO 2022/266853 A1

[Continued on next page]

MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

METHODS AND DEVICES FOR GESTURE RECOGNITION

Technical Field

[0001] Exemplary implementations described herein generally relate to methods and devices for gesture recognition.

Background

[0002] Gesture recognition through visual understanding is a fundamental task in various fields, including robotic applications. Different gestures have different hand shapes and movements. Computer vision techniques may be used to classify different gestures, e.g., through both visual appearances and hand motions. However, gesture recognition methods typically assume that the camera is static, where the hand pose in image frames results from only from hand movements. Gesture applications as in virtual reality (VR) and augmented reality (AR) also typically assume the hand is close to the camera so that the hand size is big and clear for hand/key point detection. However, this assumption is not always true in robotic fields. For robots with mobility such as AMR (autonomous mobile robots), vacuum cleaning robots, robot arms, etc., the camera (which is attached to the robot) moves in a large and free space. In such an application, difficulties in gesture recognition arise because 1) hand and camera motions are coupled together in terms of image changes, which may lead to classification error towards hand motions and 2) hand images are easily blurred due to both camera motion and far distance. These difficulties may cause failure of hand detection, key point detection, gesture classification, etc. Accordingly, approaches for gesture recognition are desirable which allow reliable gesture recognition in particular in applications with strong camera movement.

Brief Description of the Drawings

[0003] In the drawings, like reference characters generally refer to the same parts throughout the different views. The drawings are not necessarily to scale, emphasis instead generally being placed upon illustrating the principles of the invention. In the following description, various aspects are described with reference to the following drawings, in which:

FIG. 1 illustrates a gesture recognition scenario where a camera taking pictures for recognition of gestures of a human user is mounted on a moving robot.

FIG. 2 illustrates an approach for gesture recognition according to an embodiment.

FIG. 3 shows a flow diagram for gesture recognition according to an embodiment.

FIG. 4 illustrates operation and training of a generator of a generative model used for gesture recognition according to an embodiment.

FIG. 5 illustrates a close loop self-supervision according to an embodiment.

FIG. 6 shows a multi-stream gesture net with gate attention according to an embodiment.

FIG. 7 shows an example of a joint representation, a first bone representation, a second bone representation and a movement representation for a hand according to an embodiment.

FIG. 8 shows a gate attention module according to an embodiment.

FIG. 9 shows a gesture recognition device according to an embodiment.

FIG. 10 shows a flow diagram illustrating a method for recognizing a gesture according to an embodiment.

Description of Exemplary Implementations

[0004] The following detailed description refers to the accompanying drawings that show, by way of illustration, specific details and aspects of this disclosure in which the invention may be practiced. Other aspects may be utilized and structural, logical, and electrical changes may be made without departing from the scope of the invention. The various aspects of this disclosure are not necessarily mutually exclusive, as some aspects of this disclosure can be combined with one or more other aspects of this disclosure to form new aspects.

[0005] FIG. 1 illustrates a gesture recognition scenario where a camera 101 taking pictures for recognition of gestures of a human user 102 is mounted on a moving robot 103.

[0006] In a scenario of FIG.1 where the camera 101 is moving with respect to the human user 102 because of movement of the robot 103, (e.g. an AMR (autonomous mobile robot), a gesture recognition approach which is based on the assumption that the camera is static or that they are only very small movements of the camera 101 with respect to the user due to camera movement and that the hand is close to the camera may not be suitable, i.e. may deliver bad results.

[0007] It should be noted that in the present and the following examples, gestures to be recognized are hand gestures performed by a human person with his or her hand 104. However, gestures may also be performed by other body parts (e.g. a gesture may be nodding with a head) and may also be gestures performed by body parts of non-human entities, e.g. another robot.

[0008] In a scenario where the camera is static and the hand close the camera, images of the hand are clear and the hand is shown with sufficient size for hand detection and hand key point detection. An example for such a scenario is image deblur and gesture control in phones and virtually reality (VR) or augmented reality (AR) devices, e.g. VR or AR glasses. Techniques used for those applications are typically only able to handle camera jitter or shake with limited movements which are small in comparison of the hand gesture movements. In those scenarios, the camera motion is caused by hand and head shakes which are very small movements compared to the gesture movements. Furthermore, in those scenarios, hands are near to the camera and are thus easy to detect. For example, in VR glasses, the distance from hand to camera is at most the length of human arm. This distance is good for hand detection since the hand region is big in the taken images.

[0009] In a robotic application such as an AMR, however, the robot-mounted camera often strongly moves. This may cause errors in dynamic gesture recognition for techniques for VR and AR applications. In addition, the user 102 may be far away from the robot 102, e.g.,

several meters away. So, the images of the hand may be blurred, and the hand may be shown with small size which may cause failure of hand detection and hand key point detection in techniques for VR and AR applications.

[0010] In view of the above, according to various embodiments, an approach for gesture recognition is provided which allows reliable gesture recognition in case of scenarios with free camera movement, e.g., using images taken from a camera mounted on a high-speed AMR whose movement may be much larger than the movement of the hand gesture which should recognized and where the hand may be far from the camera. For example, the robot 103 includes a gesture recognition device 105 configured to operate according to this approach.

[0011] FIG. 2 illustrates an approach for gesture recognition according to an embodiment.

[0012] The approach includes processing of an input video 203 (i.e. a sequence of images showing a hand performing a gesture) by a generative model 201 and a fusion 202 of features generated from multiple gesture information streams by a multi-stream gesture map to generate a gesture recognition result 204.

[0013] The generative model 201 allows dealing with free sized hands and free camera motions. It combines motion disentangle (or compensation), image quality improvement, and gesture recognition together to boost the end-to-end performance.

[0014] The generative model 201 is trained by supervising it by a discriminator so that it is trained to enhance small sized or blurred hands for robust hand key point detection.

[0015] The two modules (generative model 201 and fusion 202) may interact with each other. They may be trained in a joint manner in an end-to-end fashion.

[0016] FIG. 3 shows a flow diagram 300 for gesture recognition according to an embodiment, for example carried out by the gesture recognition device 105.

[0017] The gesture recognition is for example applied for a robot 103 having a robot-mounted camera 101 as illustrated in FIG. 1.

[0018] According to various embodiments, the approach illustrated in FIG.2 is combined with the usage of data acquired from other sensors (i.e. in addition to the camera 101) to

improve the video data input for gesture recognition. For example, using audio or voice input can help with moving the robot camera 101 to the right direction and orientation to ensure that image data can be captured accurately for gesture recognition. Lidar or IR (infrared) sensors can also be used to accurately position the robot 103 with respect to the hand gestures to ensure accurate video data input capture of the hand 104.

[0019] Accordingly, in 301, sensor data from one or more other sensors is acquired and the robot 103 and the camera 101 are positioned such that the camera 101 can take pictures of the hand 104, i.e. can scan the hand gesture.

[0020] Once the camera 101 is accurately positioned to scan the hand gesture, the camera 101 takes a sequences of images of the hand 104 in 302.

[0021] In 303, to handle strong camera movement and small hand size, a generator of a generative model is applied, which is in this embodiment a generator of a gesture supervised recursive GAN (Generative Adversarial Network) for motion disentangle and high-quality hand image generation. The generator disentangles the hand motion (caused by the gesture) from the camera motion (caused by movement of the robot 103) and thus can handle free camera movements. The generator further increases hand resolution without loss the image quality, to boost the success rate of the hand detection and key point detection. The generator operates in a recursive manner, i.e. generates a motion-compensated (and enhanced) image for a current image of the video 203 based on the current image and the motion-compensated (and enhanced) image generated for a previous image of the video 203.

[0022] Thus, the generative model 201 generates a sequence of motion-compensated (and enhanced) images.

[0023] To perform multi-stream gesture recognition using fusion 202 of features, multiple streams are generated in 304 which represent the pose or movement of the hand. This may for example be a sequence of representations indicating joint positions, a sequence of representations indicating bone positions or a sequence of representations indicating

movement of joints. As representations, graphs (including nodes and edges) may for example be used.

[0024] The multiple representations are then processed by a gesture net (i.e. a neural network for gesture recognition) in 305, where the multiple streams are fused by the fusion 202 of features based on a gate attention. Thus, according to one embodiment, the following three sequences of representations are combined: first order (joints), spatial second order (bones) and temporal second order (motion).

[0025] In the following, the generative model used in 303 is described in detail.

[0026] FIG. 4 illustrates operation and training of a generator 400 of a generative model used for gesture recognition according to an embodiment.

[0027] The generator 400 includes two sub-generators: a motion disentangle generator 401 (also referred to as motion compensation generator) and a (high-quality) hand image generator 402. For training of the generator 400 (according to GAN training), the generative model includes a discriminator 403 which includes two sub-discriminators: an image discriminator 404 and a hand key point discriminator 405. The training of the generator 400 (as well as the discriminator 403) may be carried out by a training (computer) device. The trained generator 400 may then be stored in the gesture recognition device 105.

[0028] In FIG. 4 and in the following explanation, I_t is the current input frame (i.e. the image of the sequence of images for currently, for time index *t*, being processed by the generator 400), $g_{m,t}$ is the output of the motion disentangle generator 401 (for the current input frame), and $g_{d,t}$ is the output of the hand image generator 402.

[0029] For supervision of the motion disentangle generator 401 in training, a reference image is generated by a motion wrapping module 406 for the current input frame I_t based on the current input frame I_t and the preceding input frame I_{t-1} using SLAM (Simultaneous Localization and Mapping) pose generation, for example a SLAM system of the robot 103. This reference image for the current frame (i.e. for time index *t*) is denoted as $I_{m,t}$.

[0030] The motion disentangle generator 401 aligns frame I_t to I_0, i.e. adjusts the current input frame I_t so as to compensate the change of the camera viewpoint (i.e. to compensate the camera movement) from time 0 to time *t*. This means that it generates an image $g_{m,t}$ under the camera pose of I_0 (i.e. a reference camera viewpoint) for the current time index *t*. The motion disentangle generator 401 does this based on the current input frame I_t and based on the image generated by the generator 400 for the preceding input frame $g_{d,t-1}$. Thus, image generation by the generator 400 can be seen to be conducted recursively (or in a closed loop) wherein adjacent frames (i.e. subsequent input frames in the input sequence) are aligned with small pose changes. It should be noted that for the first input frame I_0 of the sequence of input images, where there is no preceding image, $g_{m,0}$ may simply be set to I_0 because it is already aligned with the reference camera viewpoint.

[0031] In training, the sequence of input frames I_t (t = 0, 1, ..., T) comes from a student dataset 409 of training images (or training videos).

[0032] As mentioned above, the motion disentangle generator 401 is supervised using the output of a motion wrapping module 406, denoted as $I_{m,t}$, for the current input frame I_t. For example, the motion wrapping module 406 uses the robot's SLAM system to compute the camera movement between the preceding input frame I_{t-1} and the current frame I_t and re-project the current frame I_t to the camera viewpoint from which the preceding input image I_{t-1} was taken (or generated in case of synthetic generation of training images). Each camera viewpoint is for example given by a camera coordinate system for the respective value of the time index *t*. The supervision is implemented by including a term punishing a high difference between $g_{m,t}$ and $I_{m,t}$ in the loss of the generator 400.

[0033] It should be noted that the wrapping image $I_{m,t}$ does not need to be a complete image since not every pixel on $I_{m,t}$ necessarily has a correspondence pixel on I_t. For example, due to the viewpoint difference, some pixel on $I_{m,t}$ may not be observable on I_t due to occlusion. Thus, only those valid pixels which have corresponding points on I_t are used to supervise the motion disentangle generator 401. The valid pixel mask is denoted as $M_{m,t}$ (see

the loss function given below). The motion disentangle generator 401, in contrast, generates complete images (having pixel values for all pixel locations).

[0034] The hand image generator 402 generates an image $g_{d,t}$ for the current input frame I_t with clear hand appearance for hand gesture recognition. Although hand gesture recognition does not need fine details on the hand appearance and texture, a hand detector may fail for blurred or small sized hands. Thus, the hand image generator 402 is provided to improve the image quality for robust hand detection and/or hand key point detection. In the example of FIG. 4, a hand key point detector 407 is trained on a teacher dataset of training images. The hand key point detector 407 is used in training of the generator 400 and the discriminator 403 and may also be used in 304 to extract a key point (e.g. joint) representation of a hand from a hand image.

[0035] To train the image generator 402, the hand key point detector 407 is kept fixed and the image $g_{d,t}$ generated by the hand image generator 402 is fed to the key point detector 407. The hand image generator 402 is trained such that it generates images leading to hand key point embeddings which the hand key point discriminator 405 can distinguish from hand key point embeddings generated from the images of the teacher data set as little as possible, i.e. to lower the ability of the hand key point discriminator 405 to distinguish between images generated by the hand image generator 402 and images from the teacher dataset 408 based on their hand key point embeddings generated by the hand key point embedder 407. In other words, image quality of images generated by the hand image generator 402 is supervised by minimizing a loss based on hand key point detection.

[0036] In the recursive image generation process performed by the generator 400 error accumulation and drifting may occur. Therefore, according to one embodiment, the generator 400 and the discriminator 403 are trained in a closed loop as illustrated in FIG. 5.

[0037] FIG. 5 illustrates a close loop self-supervision. Images $g_t = (g_{m,t}, g_{d,t})$ (t = 0, 1, …, T) are generated in a forward process and $g_t' = (g_{m,t}', g_{d,t}')$ are the images generated in a backward process. First, the generator 400 generates the images g_t recursively from the first

frame to the last frame, then reverses the sequence and generates frames from the last frame to the first frame. Thus, it forms a set of closed loops. The forward generated images should be the same as that from backward generated images which is reflected in the loss function for the generator 400.

[0038] The loss function L_G for the generator 400 and the loss function $L_{\mathbf{D}}$ for the discriminator 403 are according to an embodiment as follows:

$$L_{\mathbf{D}} = -\frac{1}{M}\left(\sum_{x\in\mathbf{T}} \log \mathbf{D}_{\boldsymbol{i}}(x) + \sum_{I\in\mathbf{S}} \log(\mathbf{1} - \mathbf{D}_{\boldsymbol{i}}(\mathbf{G}(I))) + \sum_{x\in\mathbf{T}} \log \mathbf{D}_{\boldsymbol{h}}(\mathbf{H}(x)) + \sum_{I\in\mathbf{S}} \log(\mathbf{1} - \mathbf{D}_{\boldsymbol{h}}(\mathbf{H}(\mathbf{G}(I))))\right)$$

$$L_{\boldsymbol{G}} = -\frac{1}{M}\left(\sum_{I\in\mathbf{S}} \log \mathbf{D}_{\boldsymbol{i}}\big(\mathbf{G}(I)\big) + \sum_{I\in\mathbf{S}} \log \mathbf{D}_{\boldsymbol{h}}(\mathbf{H}(\mathbf{G}(I))) + \sum_{I\in\mathbf{S}}\sum_{t=1}^{T} \left\|\left(g_{m,t} - I_{m,t}\right) * M_{m,t}\right\|_2\right.$$

$$\left. + \sum_{I\in\mathbf{S}}\sum_{t=1}^{T} \left\|g_{d,t} - g_{d,t}{}'\right\|_2 + \sum_{I\in\mathbf{S}}\sum_{t=1}^{T} \left\|g_{m,t} - g_{m,t}{}'\right\|_2\right)$$

[0039] As mentioned above, in the processing of a gesture net in 305, multiple streams of hand pose or hand movement representations are fused by the fusion 202 of features based on gate attention. This is explained in detail in the following.

[0040] FIG. 6 shows a multi-stream gesture net 600 with gate attention.

[0041] The gesture recognition device 105 classifies the input video 203 as one of a plurality of pre-defined gesture classes 611 by means of the multi-stream gesture net 600. In the example of FIG. 6, the multi-stream gesture net 600 is a triple-stream spatial-temporal graph convolutional network to allow robust classification (and thus gesture recognition).

[0042] The input to the multi-stream gesture net 600 are multiple streams of representations, i.e. multiple sequences of representations, wherein each representation may for example be given in graph form, i.e. be given by a graph data structure.

[0043] FIG. 7 shows an example of a joint representation 701, a first bone representation 702, a second bone representation 703 and a movement (motion) representation 704.

[0044] As mentioned above, the hand key point detector 407 may generate (in 304), for each image generated by the generator 400 the joint representation 701 (i.e. the key points are joints represented by nodes in the joint representation 701). From this, the gesture

recognition device 105 may generate the first bone representation 702 by connecting adjacent joints. The gesture recognition device 105 may generate the second bone representation 702 by defining key joints (e.g. the finger tip), connecting the key joints and connecting each remaining joint to the nearest key joint. The gesture recognition device 105 may generate the motion representation 704 by assigning, to each joint, a motion vector specifying the movement of the joint between the preceding joint representation (i.e. the joint representation generated for the preceding input frame) and the current frame.

[0045] Thus, the gesture recognition device 105 generates sequences (or streams) of representations, in this example three sequences (joint stream, bone stream and motion stream).

[0046] For input to a respective ST-GCN (Spatial Temporal Graph Convolutional Network), each sequence of representations is written as a spatial-temporal graph. For the joint representation, the input to the ST-GCN is for example the joint coordinate vectors on the graph nodes (i.e. vertices), i.e. there is a graph node for each joint and its initial features are the coordinates of the joint (so an initial feature map includes coordinates for all joints and for all input frames). These features change as a result of the processing by the ST-GCN (by combination of features of different nodes by convolution). For the motion representation, there is a graph node for each joint and its initial feature is the motion vector of the joint. For the bone stream, the initial feature of each joint is the bone vector pointing from the joint to another joint. The number of nodes is equal to the number of bones. According to one embodiment, the number of joints, bones, and motion nodes are the same for computational convenience.

[0047] The gesture net 600 includes, for each stream of representations, a chain of ST-GCNs 601 to 606 forming a chain of processing stages 607, 608. After each processing stage 607, 608, the output of the processing stage 607, 608 for each representation stream is adjusted by the outputs of the processing stage 607, 608 for the other representation streams which are supplied to the representation stream via a respective gate (e.g. gates 609, 610 for

the joint stream after the first processing stage 607) which depends on trainable parameters for the gate, wherein the parameters for different gates are independently trainable. The gates 609, 610 are part of gate attention modules (not shown in FIG. 6) which will be described in more detail below with reference to FIG. 8.

[0048] The gesture net 600 is trained by a training device (e.g. the same that trains the generator 400, the discriminator 403 and possibly also the hand key point detector 407).

[0049] FIG. 8 shows a gate attention module 800.

[0050] The gate attention module 800 serves adjusting the output of a processing stage 607, 608 for a first representation stream. This output is referred to as H_1 and it includes $C \times N \times T$ values, wherein N is the number of nodes of the (graph) representation, T is, as above, the number of input frames and C is the number of channels (i.e. the dimensionality of the feature vector assigned to each node). The gate attention module 800 adjusts the output H_1 based on the output H_2 of the processing stage for a second representation stream and the output H_3 of the processing stage for a third representation stream.

[0051] The gate attention module includes a first gate 801 and a second gate 802. For example, the first representation stream is the joint stream, the output H_1 is the output of the first processing stage 607 for the joint stream, the output H_2 is the output of the first processing stage 607 for the bone stream and the output H_3 is the output of the first processing stage 607 for the motion stream, the first gate 801 corresponds to the gate 609 and the second gate 802 corresponds to the gate 610. The gesture net 600 includes similar gate attention modules of the second representation stream and the third representation stream (wherein the indices are changed accordingly, i.e. H_1 is replaced by H_2 for the second representation stream etc.)

[0052] Each gate 801, 802 adjusts the value H_1 according to a gate attention given by an attention value. The gate attention can be directed or undirected. In the directed case, the attention value of the adjustment $i \to j$ (i.e. the adjustment of H_j by H_i) is independent from that of $j \to i$. This means that the importance of the ith representation stream to the jth

representation stream could be different from the importance of jth representation stream to the ith representation stream. In the undirected case, the ith and jth representation streams share the same importance to each other. According to various embodiments, directed gate attention is used. A gate attention module using undirected gate attention can be seen as a simplified version of the gate attention module 800 and can be deduced from it.

[0053] The gate attention module for the jth stream calculates attention values A_j^i of a directed attention map, wherein A_j^i describes the importance of each vertex (i.e. graph node) on the ith stream to the corresponding vertex on the j stream:

$$A_j^i = H_j * \omega_j^i + H_i * \omega_i^j,$$

where '*' denotes convolution, j $= 1,2,3$, $i = 1,2,3$, and $j \neq i$ (j = 1 and i = 2, 3 in FIG. 8). The ω_j^i and ω_i^j are trainable convolution kernels (i.e. specify sets of filter weights) of first to fourth convolutional filters (803, 804, 805, 806 in the example of FIG. 8) indicating the importance weight from stream i to j, and j to i, respectively. The operation of each gate is defined by a gate map

$$G_j^i = \delta(A_j^i),$$

where the $\delta(\cdot)$ is the sigmoid function. In the example of FIG. 8, for the first gate 801, i=2 and j=1 and for the second gate 802, i=3 and j=1. Each gate value $G_j^i(c,n,t)$ of the gate map depends on the channel, vertex and frame and represents how much information will be fed from stream i to j, at vertex (n,t) on the c^{th} feature component.

[0054] The attention map and the gate map is computed for every pair of streams and thus there are six attention maps and six gate maps.

[0055] Finally, the output for the kth stream (k=1 in FIG. 8) after adjustment is given by:

$$H_{k_{fusion}} = \omega_k * f^{cat}\left(G_k^i \cdot H_i, G_k^j \cdot H_j, H_k\right) + b_k,$$

where $k \neq i,j$ and $j \neq i$ (i.e. j and i are the remaining two streams, i.e. 2 and 3 in FIG. 8) f^{cat} is the concatenation operator, the "$\cdot$" is element wise multiplication, ω_k is a (trainable) convolution kernel of a fifth convolutional filter 807 and b_k is bias. The convolution kernels

have dimensions such that the output $H_{k_{fusion}}$ has again $C \times N \times T$ value and is also denoted as fusion map. The fusion map output for the *k*th stream merges feature values determined for the other two streams with the feature values determined for the *k*th stream based on the gate attention values which depend on the trainable convolutional kernels ω_k^i and ω_k^j used in the convolutional filters 803, 804, 805, 806.

[0056] The output of the last processing stage 608 is fed to one or more (e.g. fully connected) layers 612 which perform the classification (e.g. output soft values for the classes 611).

[0057] The input video has for example 20 to 50 frames. Training may for example be performed using the Dynamic Hand gesture 14/28 dataset which is split into trainings samples and testing samples (e.g. 70% to 30%).

[0058] According to various examples, in other words, a gesture recognition device is provided as illustrated in FIG. 9.

[0059] FIG. 9 shows a gesture recognition device 900 according to an embodiment.

[0060] The gesture recognition device includes an input interface 901 configured to receive a sequence of images, each image showing a body part with which a gesture is performed from a viewpoint of a camera.

[0061] Further, the gesture recognition device includes a processor 902 configured to generate a sequence of motion-compensated images from the sequence, including:
generating a motion-compensated image for an image of the sequence by compensating the movement of the camera viewpoint from a reference camera viewpoint to the viewpoint from which the image shows the body part based on

- the image and
- a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence.

[0062] The processor 902 is further configured to estimate the gesture from the sequence of motion-compensated images.

[0063] According to various embodiments, in other words, a generator (e.g. of a GAN) is used to compensate the movement of a camera taking pictures for gesture recognition. The generator has a feedback loop in the sense that it uses a motion-compensated image it has generated for an earlier point in time for generation of a motion-compensated image it generates for a current (later) time (assuming that the images are indexed by points in time).

[0064] The approach of FIG. 9 allows reliable gesture recognition with a moving camera, e.g. attached to a moving robot. It may for example be applied for hand gesture recognition which is a basic function in various fields including robotics, autonomous driving, human computer interaction, etc. In all these fields, cameras may be strongly moving. In addition, the approach of FIG. 9 can also be applied in other video analysis fields with one or more moving cameras, such as action recognition.

[0065] FIG. 10 shows a flow diagram 1000 illustrating a method for recognizing a gesture, for example performed by a gesture recognition device.

[0066] In 1001, the gesture recognition device receives a sequence of images, each image showing a body part with which a gesture is performed from a viewpoint of a camera.

[0067] In 1002, the gesture recognition device generates a sequence of motion-compensated images from the sequence. The generation of the sequence of motion-compensated images includes generating a motion-compensated image for an image of the sequence by compensating the movement of the camera viewpoint from a reference camera viewpoint to the viewpoint from which the image shows the body part based on

- the image and
- a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence.

[0068] 1003, the gesture recognition device estimates the gesture from the sequence of motion-compensated images.

[0069] The gesture recognition device may receive the sequence of images from a camera of an autonomously moving robot. The recognition device may also be part of the autonomously moving robot.

[0070] According to one embodiment, in addition or in alternative to the above, the input interface 901 is configured to receive a plurality of sequences of representations of a pose or movement of a body part and the processor 902 is configured to implement a neural network configured to determine, for each sequence of representations of the plurality of sequences of representations, a feature map and is configured to adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of the plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations to adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations, wherein the second attention map is different from the first attention map and to estimate a gesture from the sequence of motion-compensated images using the adjusted feature map values determined for the first sequence of representations and the adjusted feature map values determined for the second sequence of representations.

[0071] Further, the device 900 of FIG. 9 may also be gesture training device (with or without input interface 901) where the processor is configured to train a generator or a gate attention module or both.

[0072] The components of the gesture recognition devices and gesture recognition training devices may be implemented by one or more processors. A "processor" may be understood as any kind of a logic implementing entity, which may be special purpose circuitry or a processor executing software stored in a memory, firmware, or any combination thereof. Thus a "processor" may be a hard-wired logic processor or a programmable logic processor such as a programmable processor, e.g. a microprocessor. A "processor" may also be a processor executing software, e.g. any kind of computer program. Any other kind of implementation of

the respective functions which will be described in more detail below may also be understood as a "processor". The communication device may for example be at least partially implemented by a transceiver which may for example be at least partially implemented by a modem (e.g. an LTE modem), a baseband processor or other transceiver components or also by an application processor. The communication device may for example be a communication terminal as such and may include typical communication terminal devices such as a transceiver (including e.g. a baseband processor, one or more filters, transmit chains, receive chains, amplifiers etc.), an antenna, a subscriber identity module, an application processor, a memory etc.

[0073] The following examples pertain to further exemplary implementations.

[0074] Example 1 is a gesture recognition device including an input interface configured to receive a sequence of images, each image showing a body part with which a gesture is performed from a viewpoint of a camera; and a processor configured to generate a sequence of motion-compensated images from the sequence, including generating a motion-compensated image for an image of the sequence by compensating the movement of the camera viewpoint from a reference camera viewpoint to the viewpoint from which the image shows the body part based on the image and a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence; and estimate the gesture from the sequence of motion-compensated images.

[0075] Example 2 is the gesture recognition device of Example 1, wherein the processor is configured to implement a generator of a generative adversarial network and is configured to generate the motion-compensated image from the sequence of motion-compensated images by means of the generator.

[0076] Example 3 is the gesture recognition device of Example 1 or 2, wherein the processor is configured to generate the motion-compensated image by generating an intermediate image, wherein in the intermediate image the movement of the camera from a reference camera viewpoint to the viewpoint from which the image shows the body part is

compensated, and wherein the processor is configured to and generate the motion-compensated image by increasing the image quality of the intermediate image.

[0077] Example 4 is the gesture recognition device of Example 3, wherein increasing the image quality of the intermediate image includes deblurring the body part in the intermediate image.

[0078] Example 5 is the gesture recognition device of any one of Examples 1 to 4, wherein the processor is further configured to generate, from the a sequence of motion-compensated images, a plurality of sequences of representations of a pose or movement of a body part; implement a neural network configured to determine, for each sequence of representations of the plurality of sequences of representations, a feature map; and a gate attention module configured to adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of the plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations; to adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations; and to estimate a gesture from the sequence of motion-compensated images using the adjusted feature map values determined for the first sequence of representations and the adjusted feature map values determined for the second sequence of representations.

[0079] Example 6 is the gesture recognition device of Example 5, wherein the second attention map is different from the first attention map.

[0080] Example 7 is a method for recognizing a gesture including receiving a sequence of images, each image showing a body part with which a gesture is performed from a viewpoint of a camera; generating a sequence of motion-compensated images from the sequence, including: generating a motion-compensated image for an image of the sequence by compensating the movement of the camera viewpoint from a reference camera viewpoint to the viewpoint from which the image shows the body part based on the image and a motion-

compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence; and estimating the gesture from the sequence of motion-compensated images.

[0081] Example 8 is a gesture recognition training device including: a processor configured to train a generator for generating a motion-compensate image for an image of a sequence of images taken by a camera, each image showing a body part with which a gesture is performed from a viewpoint of a camera, by compensating the movement of the camera from a reference camera viewpoint to the viewpoint from which the image shows the body part based on the image and based on a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence by adjusting the generator to lower the ability of a discriminator to detect that a motion-compensated image generated by the generator has been generated by the generator.

[0082] Example 9 is the gesture recognition training device of Example 8, wherein the processor is further configured to train the discriminator to distinguish between motion-compensated images generated by the generator and images from a reference image data set.

[0083] Example 10 is the gesture recognition training device of Example 9, wherein the processor is configured to train the discriminator to distinguish a motion-compensated image generated by the generator and an image from the reference image data set based on key point embeddings generated for the motion-compensated image and the image from the reference image data set.

[0084] Example 11 is the gesture recognition training device of Example 8 or 9, wherein the generator is configured to generate the motion-compensated image by generating an intermediate image in which the movement of the camera from a reference camera viewpoint to the viewpoint from which the image shows the body part is compensated and generating the motion-compensated image by increasing the image quality of the intermediate image and wherein the processor is configured to train the generator to reduce the difference between the intermediate image and a reference intermediate image generated from the motion-

compensated image by motion compensation performed based on a determination of a camera viewpoint difference from the image and the preceding image.

[0085] Example 12 is the gesture recognition training device of Example 11, wherein the processor is configured to determine the camera viewpoint difference by Simultaneous Localization and Mapping.

[0086] Example 13 is a method for training a generator including adjusting a generator for generating a motion-compensate image for an image of a sequence of images taken by a camera, each image showing a body part with which a gesture is performed from a viewpoint of a camera, by compensating the movement of the camera from a reference camera viewpoint to the viewpoint from which the image shows the body part based on the image and based on a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence, wherein the generator is adjusted to lower the ability of a discriminator to detect that a motion-compensated image generated by the generator has been generated by the generator.

[0087] Example 14 is a gesture recognition device including: an input interface configured to receive a plurality of sequences of representations of a pose or movement of a body part; and a processor configured to implement a neural network configured to determine, for each sequence of representations of the plurality of sequences of representations, a feature map; and adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of the plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations; adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations, wherein the second attention map is different from the first attention map; and estimate a gesture from the sequence of motion-compensated images using the adjusted feature map values determined for the first sequence of representations and the adjusted feature map values determined for the second sequence of representations.

[0088] Example 15 is the gesture recognition device of Example 14, wherein the processor is configured to determine the first attention map and the second attention map.

[0089] Example 16 is the gesture recognition device of Example 14 or 15, wherein the processor is configured to determine the first attention map depending on a first set of neural network parameters and the second attention map depending on a second set of neural network parameters, wherein the neural network parameters are independently trainable.

[0090] Example 17 is the gesture recognition device of Example 16, wherein the neural network is a convolutional neural network and the neural network parameters of the first set of neural network parameters and of the second set of neural network parameters are filter weights.

[0091] Example 18 is the gesture recognition device of any one of Examples 14 to 17, wherein the neural network is configured to determine the feature map by processing an input feature map.

[0092] Example 19 is the gesture recognition device of any one of Examples 14 to 18, wherein the processor is configured to determine the first attention map and the second attention map based on the input feature map.

[0093] Example 20 is the gesture recognition device of any one of Examples 14 to 19, wherein the processor is configured to determine the first attention map from the feature map values determined for the first sequence of representations and to determine the second attention map from the feature map values determined for the second sequence of representations.

[0094] Example 21 is the gesture recognition device of any one of Examples 14 to 20, wherein each sequence of representation has a plurality of nodes and the feature map for the sequence of representations has a vector of feature values for each node.

[0095] Example 22 is the gesture recognition device of any one of Examples 14 to 21, wherein each representation has nodes and the feature map for the sequence of representations has a vector of feature values for each node for each representation.

[0096] Example 23 is the gesture recognition device of any one of Examples 14 to 22, wherein the neural network includes a sequence of neural network stages beginning with a first neural network stage and ending with a last neural network stage, wherein the first neural network stage is configured to receive as input, for each sequence of representations, an initial feature map and each neural network stage is configured to generate, for each sequence of representations, a feature map based on a feature map it receives as input and each neural network stage except the last is configured to provide the feature map it has generated as input to the following neural network stage; and the processor is configured to, for each neural network stage adjust, in accordance with a respective first attention map, feature map values determined by the neural network stage for the first sequence of representations, by feature map values determined for the second sequence of representations of the plurality of sequences of representations; and to adjust, in accordance with a respective second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations, wherein the second attention map is different from the first attention map.

[0097] Example 24 is the gesture recognition device of Example 23, wherein for at least one sequence of representations, the initial feature map includes joint positions of the body part.

[0098] Example 25 is the gesture recognition device of Example 23 or 24, wherein for at least one sequence of representations, the initial feature map includes movement vectors of joints of the body part specifying movement of joints over the sequence of representations.

[0099] Example 26 is a method for recognizing a gesture including: receiving a plurality of sequences of representations of a pose or movement of a body part; determining, by a neural network, for each sequence of representations of the plurality of sequences of representations, a feature map; adjusting, in accordance with a first attention map, feature map values determined for a first sequence of representations of the plurality of sequences of representations, by feature map values determined for a second sequence of representations of

the plurality of sequences of representations; adjusting, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations, wherein the second attention map is different from the first attention map; and estimating a gesture from the sequence of motion-compensated images using the adjusted feature map values determined for the first sequence of representations and the adjusted feature map values determined for the second sequence of representations.

[00100] Example 27 is a gesture recognition training device including: a processor configured to train a gate attention module to adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of a plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations and adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations by adjusting a first set of neural network parameters and a second set of neural network parameters, wherein the gate attention module determines the first attention map from the feature map values determined for the first sequence of representations depending on the first set of neural network parameters and the gate attention module determines the second attention map from the feature map values determined for the second sequence of representations depending on the second set of neural network parameters.

[00101] Example 28 is a method for training a gate attention module for a gesture recognition device, including:training the gate attention module to adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of a plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations and adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of

representations by adjusting a first set of neural network parameters and a second set of neural network parameters, wherein the gate attention module determines the first attention map from the feature map values determined for the first sequence of representations depending on the first set of neural network parameters and the gate attention module determines the second attention map from the feature map values determined for the second sequence of representations depending on the second set of neural network parameters.

[00102] It should be noted that one or more of the features of any of the examples above may be combined with any one of the other examples.

[00103] While specific aspects have been described, it should be understood by those skilled in the art that various changes in form and detail may be made therein without departing from the spirit and scope of the aspects of this disclosure as defined by the appended claims. The scope is thus indicated by the appended claims and all changes which come within the meaning and range of equivalency of the claims are therefore intended to be embraced.

Claims

1. A gesture recognition device comprising:

an input interface configured to receive a sequence of images, each image showing a body part with which a gesture is performed from a viewpoint of a camera; and

a processor configured to

generate a sequence of motion-compensated images from the sequence, comprising:

generating a motion-compensated image for an image of the sequence by compensating the movement of the camera viewpoint from a reference camera viewpoint to the viewpoint from which the image shows the body part based on

the image and

a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence; and

estimate the gesture from the sequence of motion-compensated images.

2. The gesture recognition device of claim 1, wherein the processor is configured to implement a generator of a generative adversarial network and is configured to generate the motion-compensated image from the sequence of motion-compensated images by means of the generator.

3. The gesture recognition device of claim 1, wherein the processor is configured to generate the motion-compensated image by generating an intermediate image, wherein in the intermediate image the movement of the camera from a reference camera viewpoint to the viewpoint from which the image shows the body part is compensated,

and wherein the processor is configured to generate the motion-compensated image by increasing the image quality of the intermediate image.

4. The gesture recognition device of claim 3, wherein increasing the image quality of the intermediate image comprises deblurring the body part in the intermediate image.

5. The gesture recognition device of claim 1, wherein the processor is further configured to generate, from the a sequence of motion-compensated images, a plurality of sequences of representations of a pose or movement of a body part;

implement

a neural network configured to determine, for each sequence of representations of the plurality of sequences of representations, a feature map; and

a gate attention module configured to adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of the plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations and to adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations; and

estimate a gesture from the sequence of motion-compensated images using the adjusted feature map values determined for the first sequence of representations and the adjusted feature map values determined for the second sequence of representations.

6. The gesture recognition device of claim 5, wherein the second attention map is different from the first attention map.

7. A method for recognizing a gesture comprising

receiving a sequence of images, each image showing a body part with which a gesture is performed from a viewpoint of a camera;

generating a sequence of motion-compensated images from the sequence, comprising:

generating a motion-compensated image for an image of the sequence by compensating the movement of the camera viewpoint from a reference camera viewpoint to the viewpoint from which the image shows the body part based on

the image and

a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence; and

estimating the gesture from the sequence of motion-compensated images.

8. A gesture recognition training device comprising:

a processor configured to train a generator for generating a motion-compensate image for an image of a sequence of images taken by a camera, each image showing a body part with which a gesture is performed from a viewpoint of a camera, by compensating the movement of the camera from a reference camera viewpoint to the viewpoint from which the image shows the body part based on the image and based on a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence by adjusting the generator to lower the ability of a discriminator to detect that a motion-compensated image generated by the generator has been generated by the generator.

9. The gesture recognition training device of claim 8, wherein the processor is further configured to train the discriminator to distinguish between motion-compensated images generated by the generator and images from a reference image data set.

10. The gesture recognition training device of claim 9, wherein the processor is configured to train the discriminator to distinguish a motion-compensated image generated by the generator and an image from the reference image data set based on key point embeddings generated for the motion-compensated image and the image from the reference image data set.

11. The gesture recognition training device of claim 8, wherein the generator is configured to generate the motion-compensated image by generating an intermediate image in which the movement of the camera from a reference camera viewpoint to the viewpoint from which the image shows the body part is compensated and generating the motion-compensated image by increasing the image quality of the intermediate image and wherein the processor is configured to train the generator to reduce the difference between the intermediate image and a reference intermediate image generated from the motion-compensated image by motion compensation performed based on a determination of a camera viewpoint difference from the image and the preceding image.

12. The gesture recognition training device of claim 11, wherein the processor is configured to determine the camera viewpoint difference by Simultaneous Localization and Mapping.

13. A method for training a generator comprising:
adjusting a generator for generating a motion-compensate image for an image of a sequence of images taken by a camera, each image showing a body part with which a gesture is performed from a viewpoint of a camera, by compensating the movement of the camera from a reference camera viewpoint to the viewpoint from which the image

shows the body part based on the image and based on a motion-compensated image of the sequence generated for a preceding image of the sequence which precedes the image in the sequence, wherein the generator is adjusted to lower the ability of a discriminator to detect that a motion-compensated image generated by the generator has been generated by the generator.

14. A gesture recognition device comprising:

an input interface configured to receive a plurality of sequences of representations of a pose or movement of a body part; and

a processor configured to

implement a neural network configured to determine, for each sequence of representations of the plurality of sequences of representations, a feature map; and

adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of the plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations;

adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations, wherein the second attention map is different from the first attention map; and

estimate a gesture from the sequence of motion-compensated images using the adjusted feature map values determined for the first sequence of representations and the adjusted feature map values determined for the second sequence of representations.

15. The gesture recognition device of claim 14, wherein the processor is configured to determine the first attention map and the second attention map.

16. The gesture recognition device of claim 14, wherein the processor is configured to determine the first attention map depending on a first set of neural network parameters and the second attention map depending on a second set of neural network parameters, wherein the neural network parameters are independently trainable.

17. The gesture recognition device of claim 16, wherein the neural network is a convolutional neural network and the neural network parameters of the first set of neural network parameters and of the second set of neural network parameters are filter weights.

18. The gesture recognition device of claim 14, wherein the neural network is configured to determine the feature map by processing an input feature map.

19. The gesture recognition device of claim 14, wherein the processor is configured to determine the first attention map and the second attention map based on the input feature map.

20. The gesture recognition device of claim 14, wherein the processor is configured to determine the first attention map from the feature map values determined for the first sequence of representations and to determine the second attention map from the feature map values determined for the second sequence of representations.

21. The gesture recognition device of claim 14, wherein each sequence of representation has a plurality of nodes and the feature map for the sequence of representations has a vector of feature values for each node.

22. The gesture recognition device of claim 14, wherein each representation has nodes and the feature map for the sequence of representations has a vector of feature values for each node for each representation.

23. A method for recognizing a gesture comprising:

receiving a plurality of sequences of representations of a pose or movement of a body part;

determining, by a neural network, for each sequence of representations of the plurality of sequences of representations, a feature map;

adjusting, in accordance with a first attention map, feature map values determined for a first sequence of representations of the plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations;

adjusting, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations, wherein the second attention map is different from the first attention map; and

estimating a gesture from the sequence of motion-compensated images using the adjusted feature map values determined for the first sequence of representations and the adjusted feature map values determined for the second sequence of representations.

24. A gesture recognition training device comprising:

a processor configured to train a gate attention module to

adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of a plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations and

adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations;

by adjusting a first set of neural network parameters and a second set of neural network parameters, wherein

the gate attention module determines the first attention map from the feature map values determined for the first sequence of representations depending on the first set of neural network parameters; and

the gate attention module determines the second attention map from the feature map values determined for the second sequence of representations depending on the second set of neural network parameters.

25. A method for training a gate attention module for a gesture recognition device, comprising:

training the gate attention module to

adjust, in accordance with a first attention map, feature map values determined for a first sequence of representations of a plurality of sequences of representations, by feature map values determined for a second sequence of representations of the plurality of sequences of representations and

adjust, in accordance with a second attention map, feature map values determined for the second sequence of representations, by feature map values determined for the first sequence of representations;

by adjusting a first set of neural network parameters and a second set of neural network parameters, wherein

the gate attention module determines the first attention map from the feature map values determined for the first sequence of representations depending on the first set of neural network parameters; and

the gate attention module determines the second attention map from the feature map values determined for the second sequence of representations depending on the second set of neural network parameters.

FIG 1

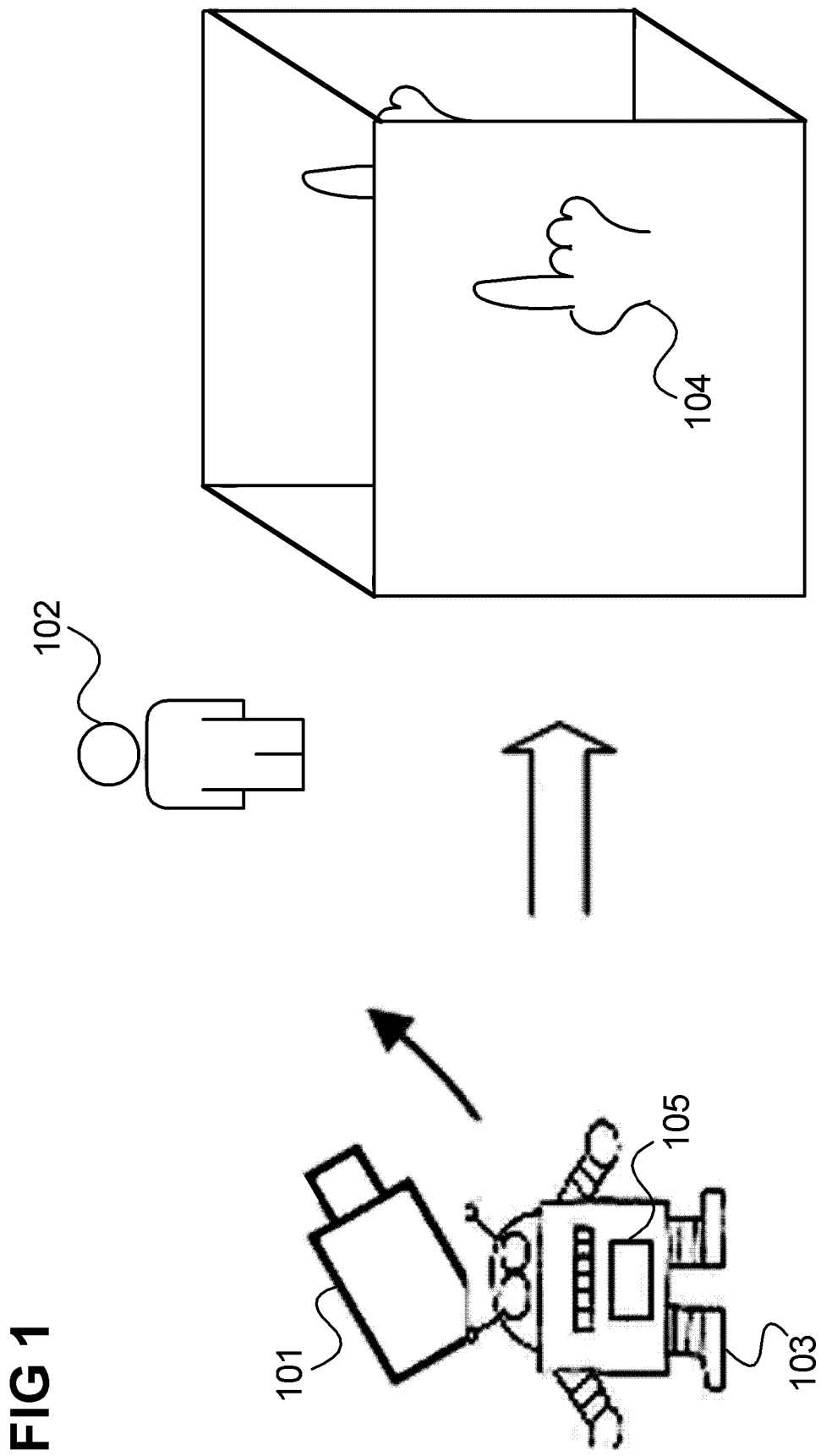


FIG 2

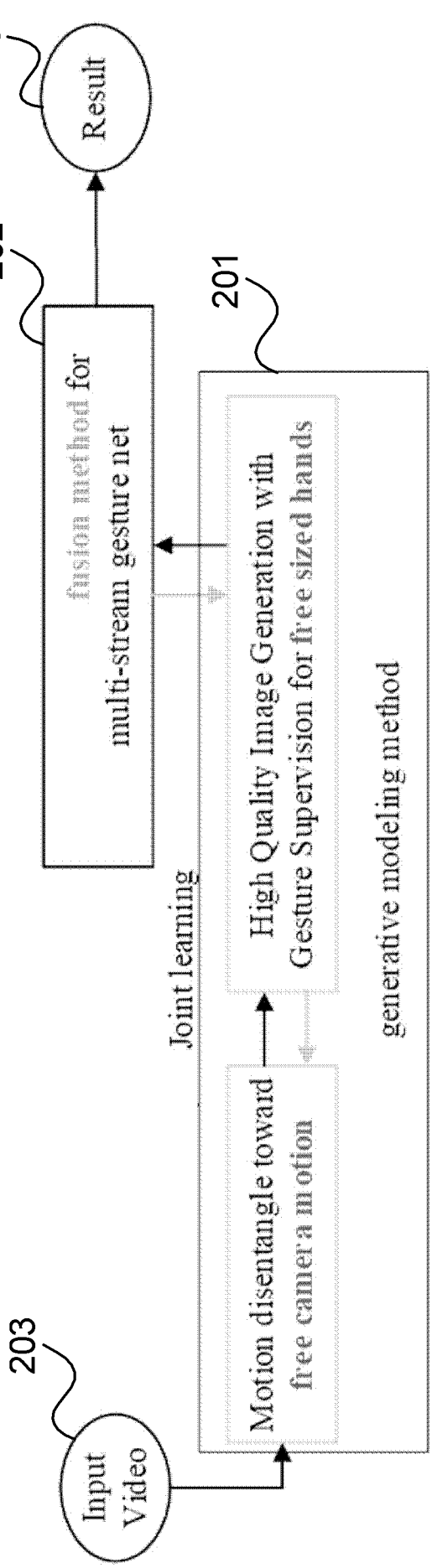

204
Result
202
fusion method for
multi-stream gesture net
201
Joint learning
High Quality Image Generation with
Gesture Supervision for free sized hands
Motion disentangle toward
free camera motion
generative modeling method
203
Input
Video

FIG 3

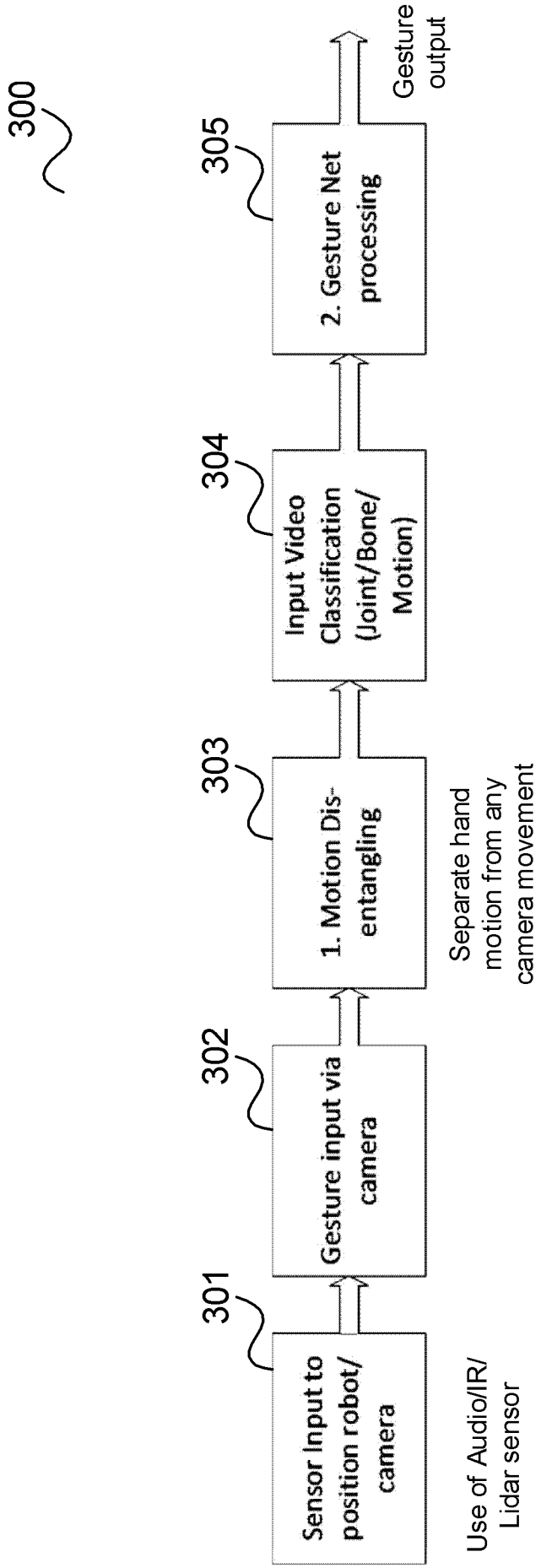

300
301
Sensor Input to position robot/ camera
Use of Audio/IR/ Lidar sensor
302
Gesture input via camera
303
1. Motion Dis-entangling
Separate hand motion from any camera movement
304
Input Video Classification (Joint/Bone/ Motion)
305
2. Gesture Net processing
Gesture output

FIG 4

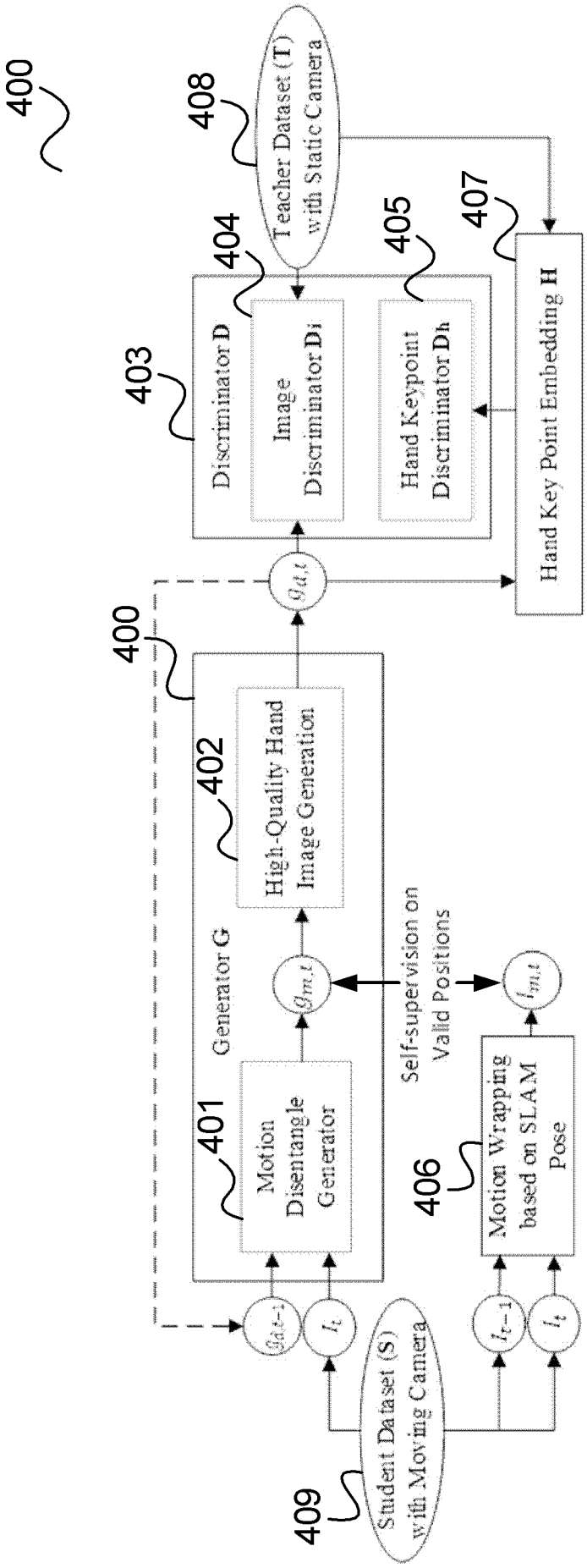
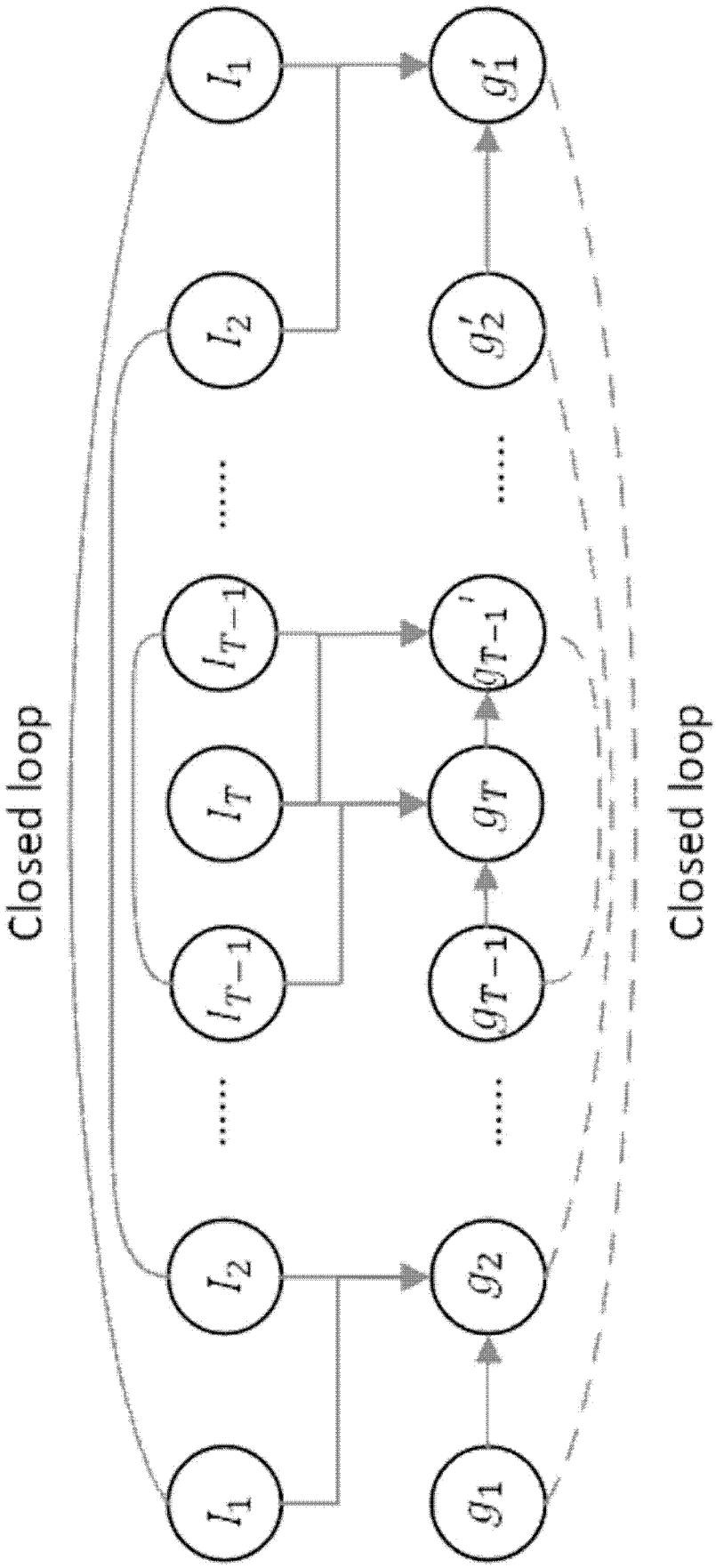

400
408
Teacher Dataset (T) with Static Camera
404
405
407
403
Discriminator D
Image Discriminator Di
Hand Keypoint Discriminator Dh
Hand Key Point Embedding H
400
402
High-Quality Hand Image Generation
Generator G
401
Motion Disentangle Generator
Self-supervision on Valid Positions
406
Motion Wrapping based on SLAM Pose
409
Student Dataset (S) with Moving Camera

FIG 5


Closed loop
Closed loop
I_1
I_2
I_{T-1}
I_T
I_{T-1}
I_2
I_1
g_1'
g_2'
g_{T-1}'
g_T
g_{T-1}
g_2
g_1

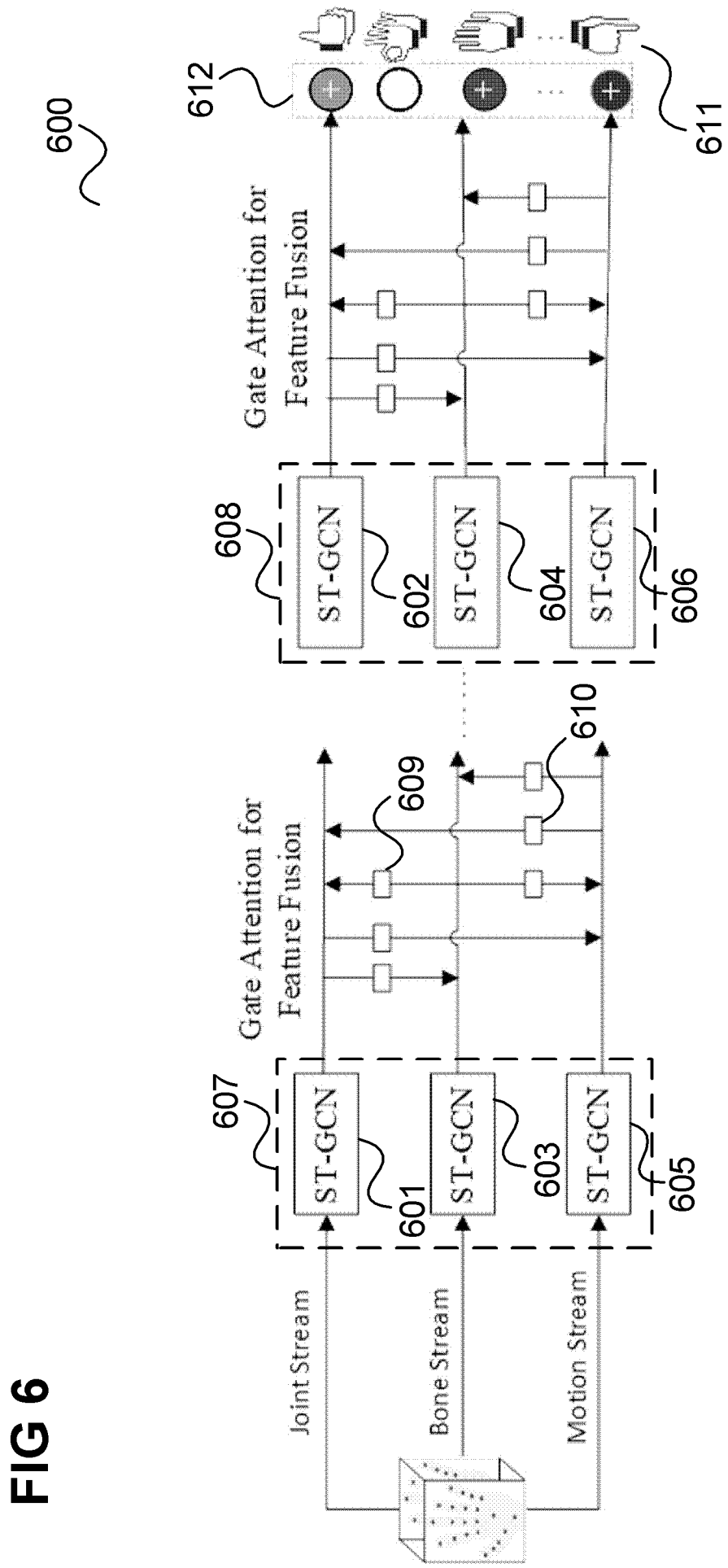

600
612
611
Gate Attention for Feature Fusion
608
ST-GCN
602
ST-GCN
604
ST-GCN
606
610
609
Gate Attention for Feature Fusion
607
ST-GCN
601
ST-GCN
603
ST-GCN
605
Joint Stream
Bone Stream
Motion Stream

FIG 6

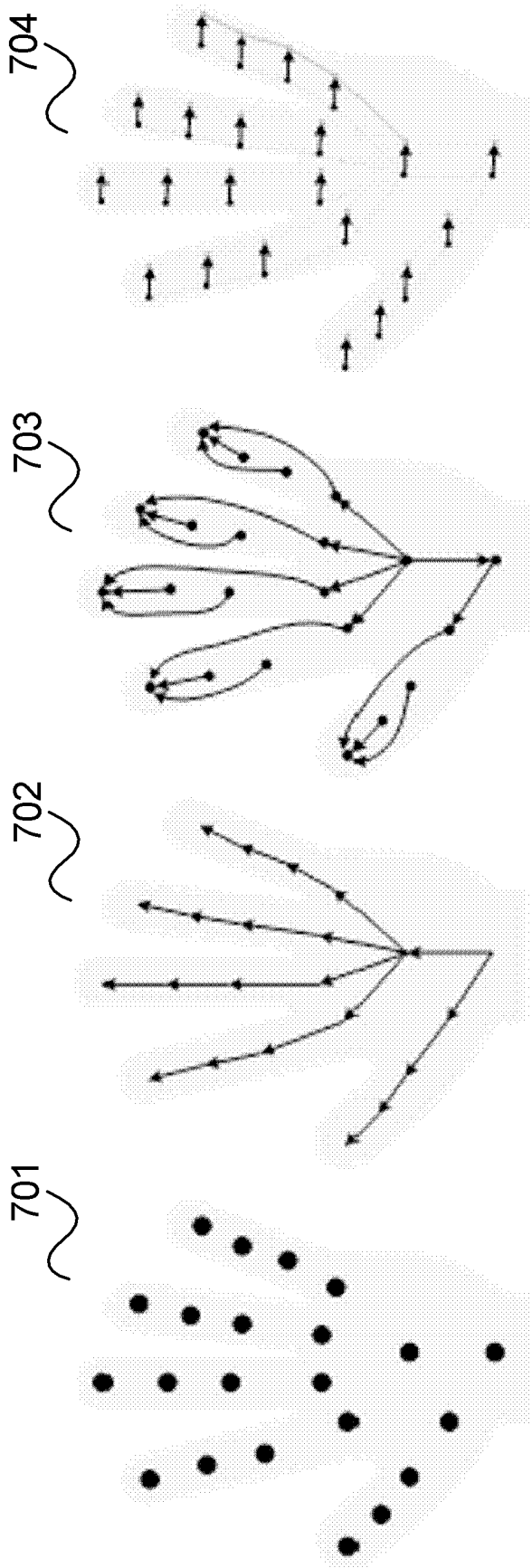
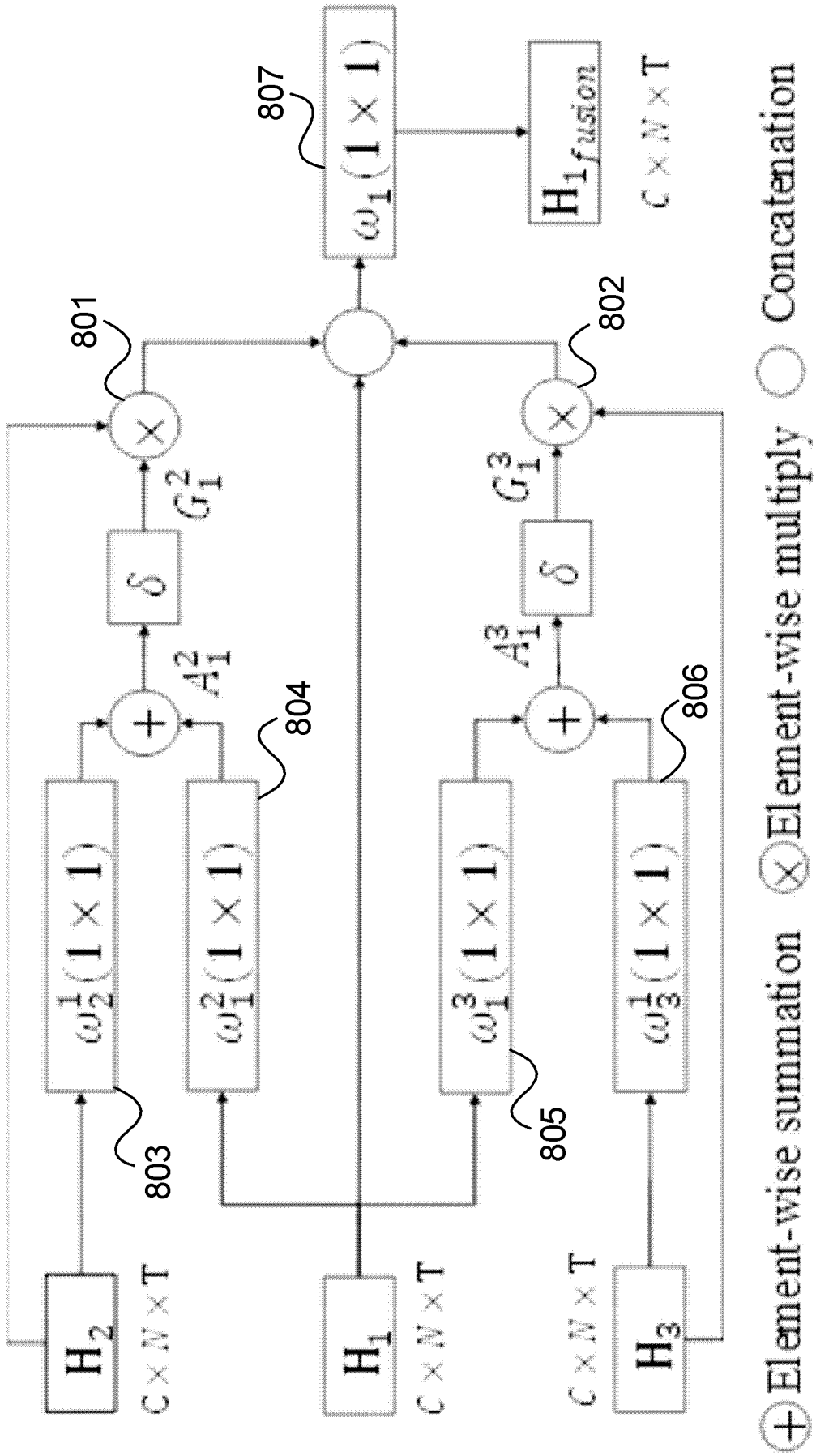


FIG 7

FIG 8


H2
C × N × T
H1
C × N × T
H3
C × N × T
ω2 1 (1 × 1)
ω1 2 (1 × 1)
ω1 3 (1 × 1)
ω3 1 (1 × 1)
803
804
805
806
801
802
807
A1 2
A1 3
δ
δ
G1 2
G1 3
ω1 (1 × 1)
H1fusion
C × N × T
Element-wise summation
Element-wise multiply
Concatenation

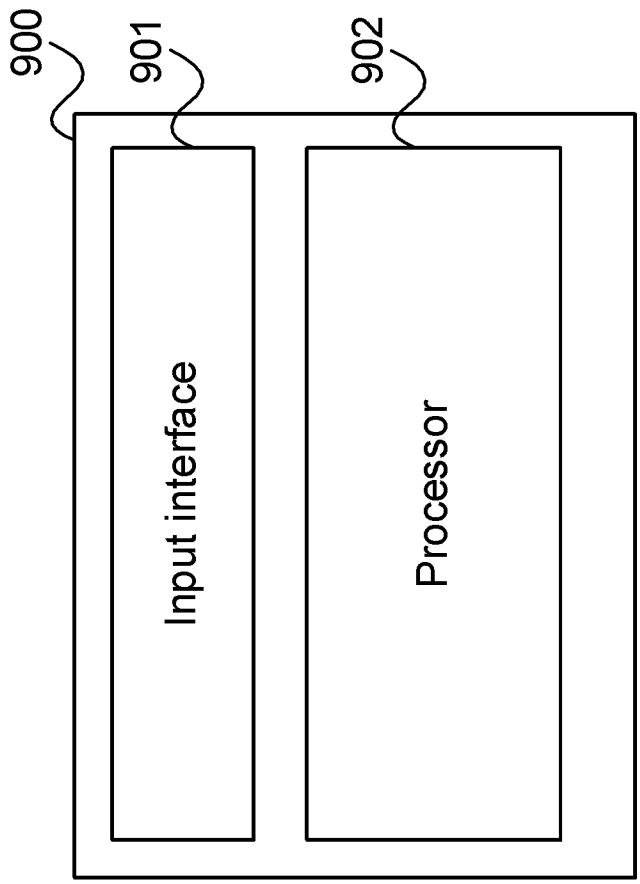
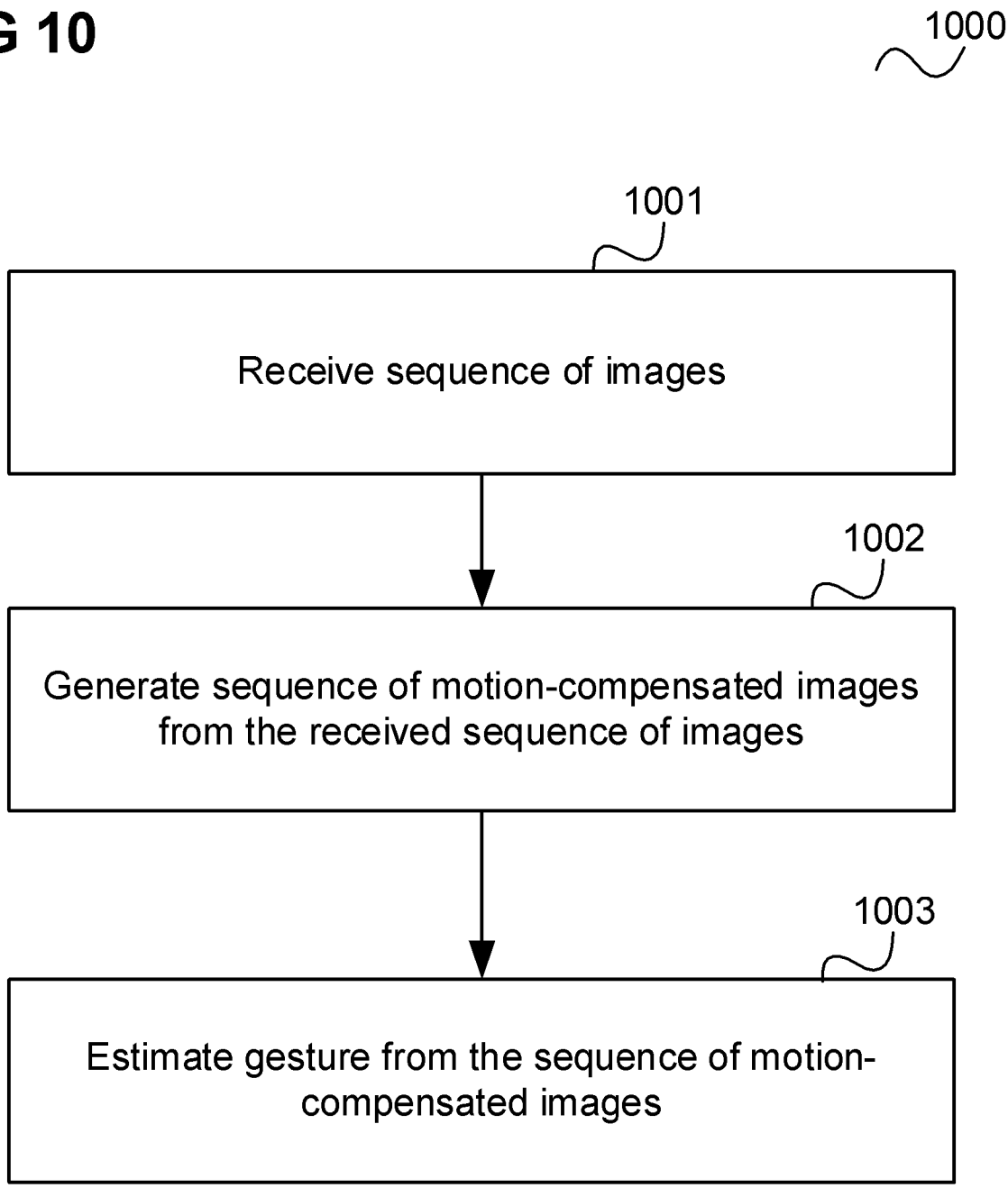


FIG 9

FIG 10


1000
1001
Receive sequence of images
1002
Generate sequence of motion-compensated images from the received sequence of images
1003
Estimate gesture from the sequence of motion-compensated images

INTERNATIONAL SEARCH REPORT

International application No.
PCT/CN2021/101611

A. CLASSIFICATION OF SUBJECT MATTER

G06K 9/00(2022.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)

G06K; G06F; H04N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

CNKI, CNPAT, WPI, EPODOC, IEEE: gesture, hand, pose, recognize, sequence, image, picture, body, part, viewpoint, reference, camera, motion, compensate, movement, preceding, estimate, adjust, generator, discriminator, neural, network, feature, map, attention, gate

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	US 2015262002 A1 (OMRON CORPORATION) 17 September 2015 (2015-09-17) description, paragraphs 0039-0116, figures 1-13	1-13
A	CN 111814626 A (SOUTH-CENTRAL UNIVERSITY FOR NATIONALITIES) 23 October 2020 (2020-10-23) claims 1-6	14-25
A	CN 111753764 A (JINAN INSPUR HIGH-TECH. INVESTMENT DEV. CO.) 09 October 2020 (2020-10-09) the whole document	1-25
A	CN 111191627 A (ZHEJIANG UNIVERSITY OF TECHNOLOGY) 22 May 2020 (2020-05-22) the whole document	1-25
A	US 2020372246 A1 (MAGIC LEAP, INC.) 26 November 2020 (2020-11-26) the whole document	1-25

☐ Further documents are listed in the continuation of Box C. ☑ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
11 March 2022	**22 March 2022**
Name and mailing address of the ISA/CN **National Intellectual Property Administration, PRC** **6, Xitucheng Rd., Jimen Bridge, Haidian District, Beijing** **100088, China** Facsimile No. **(86-10)62019451**	Authorized officer **ZHANG,Wen** Telephone No. **86-(10)-53961314**

INTERNATIONAL SEARCH REPORT
Information on patent family members

International application No.
PCT/CN2021/101611

Patent document cited in search report			Publication date (day/month/year)	Patent family member(s)			Publication date (day/month/year)
US	2015262002	A1	17 September 2015	CN	104914988	A	16 September 2015
				KR	20150107597	A	23 September 2015
				JP	2015176253	A	05 October 2015
CN	111814626	A	23 October 2020		None		
CN	111753764	A	09 October 2020		None		
CN	111191627	A	22 May 2020		None		
US	2020372246	A1	26 November 2020	WO	2020236993	A1	26 November 2020